Experimental Protocol – XCS224U – Fine-tuning vs prompting in adaptation of language model for structured text generation

Georgii Shushuev
AI Professional Program
Stanford Online Course
XCS224U
shyshyev@gmail.com

Sergey Grebenkin
AI Professional Program
Stanford Online Course
XCS224U
sergey.grebenkin@gmail.com

1 Hypotheses

In this study, we aim to explore the effectiveness of different approaches for converting free-text into structured formats. Our investigation is guided by several key hypotheses that address the comparative performance of fine-tuning and prompting methods. These hypotheses are designed to evaluate whether traditional fine-tuning techniques or innovative prompt engineering can better meet the demands of structured output tasks in natural language processing.

We propose to investigate the following hypotheses: Fine-tuning BERT-like models on well-curated datasets could yield a better performance than prompting a powerful LLM. This suggests that traditional fine-tuning of models such as BERT, when provided with high-quality, task-specific data, will outperform the use of prompts with large language models like GPT-4 for structured output tasks.

Additionally, we hypothesize that fine-tuning an LLM combined with simple prompting achieves higher quality than complex prompting without fine-tuning. This implies that a hybrid approach, involving both fine-tuning a smaller LLM (e.g., a fine-tuned version of a smaller model like LLaMA) and using straightforward prompts, will result in better performance compared to relying solely on intricate prompt engineering with no fine-tuning.

Furthermore, it is hypothesized that it is always possible to engineer a prompt that outperforms any fine-tuned model for specific tasks. This explores the potential of prompt engineering, positing that with sufficient effort and understanding of the task, one can create prompts that enable large models to achieve superior performance compared to fine-tuned models, especially for well-defined tasks. Some studies like (3) show that this could be an effective approach to narrow the solution to a simplier model.

The overarching goal of these hypotheses is to assess whether BERT-like models or smaller LLMs can be optimized to match the performance of commercial GPT models, particularly in terms of accuracy and reliability for structured output tasks.

2 Data

In this study, we will utilize several datasets to evaluate the performance of fine-tuning and prompting methods for structured output tasks. The selected datasets cover a range of text classification tasks, including multilabel classification, to provide a comprehensive assessment of the models' capabilities.

For multilabel classification tasks, we will select datasets that require the assignment of multiple labels to each text instance. A good source for such datasets is Kaggle, where various multilabel classification datasets are available. One example is the ArXiv Paper Abstracts dataset (1). This dataset includes abstracts of academic papers with multiple labels corresponding to different research topics, providing a challenging testbed for our models. Another one is (2) to classify the research papers based on Abstract and Title.

In real-world production environments, labeled datasets are often scarce and expensive to create. This poses a significant challenge for training and fine-tuning models. To address this issue, we propose leveraging large language models (LLMs) with well-crafted prompts to generate labeled data. By using prompt-based methods, we annotate unlabeled data efficiently, reducing the need for extensive manual labeling efforts. This approach not only saves time and resources but also enables the rapid adaptation of models to new tasks and domains.

Although our focus is on structured output tasks, we use classification datasets to illustrate our approach. For example, sentiment analysis is a classification task that can be effectively solved by fine-tuning BERT. When dealing with non-standard classification tasks where labeled datasets are unavailable, a well-designed prompt for a powerful LLM (e.g., GPT-4) can be used to annotate large amounts of text, which can then be used to fine-tune a model.

The key idea is that tasks not requiring free-form text generation—such as multilabel-multiclass classification, named entity recognition (NER), information extraction, and even converting free text to a structured JSON format—can be addressed by fine-tuning models that do not generate text. While it might seem convenient to solve the task of converting text to a structured JSON format by describing the transformation rules in a prompt, essentially, we are asking the LLM to perform multiple classification/NER tasks.

Structured output tasks often involve transforming unstructured text into a predefined format, such as JSON. This transformation can be broken down into several smaller tasks, such as identifying entities, classifying text segments, and assigning labels. Multilabel classification is a specific type of structured output task where each text instance can belong to multiple categories simultaneously. By using multilabel classification datasets, we can simulate the complexity of real-world structured output tasks and evaluate how well our models can handle multiple labels and categories.

By combining these datasets and methodologies, we aim to thoroughly evaluate the effectiveness of fine-tuning and prompting techniques in various text classification scenarios. This will help us understand the trade-offs and potential benefits of each approach in both research and practical applications.

3 Metrics

In this study, we will use standard evaluation metrics for multi-label classification to assess the performance of our models. These metrics are well-established in the field of natural language processing and provide a quantitative basis for comparing different approaches. The primary metrics we will use include accuracy, which measures the proportion of correctly predicted labels to the total number of labels. In multi-label classification, this metric can be adapted to account for multiple labels per instance.

Precision is the ratio of true positive predictions to the total number of positive predictions. It indicates the accuracy of the positive predictions made by the model. Recall, also known as sensitivity, is the ratio of true positive predictions to the total number of actual positives. It measures the model's ability to identify all relevant instances. The F1-score is the harmonic mean of precision and recall. It provides a single metric that balances both precision and recall, making it useful for evaluating models where there is an uneven class distribution.

A crucial metric for our specific task is the Exact Match Ratio (also known as Subset Accuracy). This metric measures the proportion of instances where the predicted set of labels exactly matches the true set of labels. For our purposes, it is essential that the entire set of labels for each instance is predicted correctly. For example, if an instance requires the prediction of three labels, the model's performance is only considered successful if all three labels are accurately predicted. If only two out of three labels are predicted correctly, it is counted as an error. This strict requirement emphasizes the importance of exact match ratio in our evaluation, as it directly reflects the model's ability to fully and correctly understand and classify each instance.

In addition to these metrics, we will also perform a detailed analysis of the model's performance on individual labels. This involves tracking metrics separately for each class and examining confusion matrices to understand common misclassifications. Confusion matrices provide a visual representation of the model's performance, showing the true versus predicted labels and highlighting areas where the model may be confusing one label for another.

When modifying prompts for specific labels, it is crucial to monitor these metrics both overall and on a per-class basis. This allows us to identify the impact of prompt changes on individual labels and ensure that improvements in one area do not come at the expense of performance in others.

By using these metrics, we aim to provide a comprehensive evaluation of the models' performance, enabling us to compare the effectiveness of fine-tuning and prompting methods in multi-label classification tasks.

4 Models

In this study, we will utilize a range of models to serve as baselines and to focus our investigation on their performance in structured output tasks. We aim to use popular and well-established architectures, ensuring that our findings are robust and applicable to real-world scenarios. The models we plan to use include: BERT, RoBERTa, ELECTRA and others

Firstly, we will employ simple models such as logistic regression combined with TF-IDF (Term Frequency-Inverse Document Frequency) features. This approach provides a straightforward baseline, allowing us to compare the performance of more complex models against a simple, interpretable method.

Secondly, we will use the mentioned BERT-like models. As it has been already shown by in (4) and (5) these models have proven highly effective for various NLP tasks, including text classification and information extraction. Fine-tuning BERT on our specific datasets could allow to leverage its pre-trained knowledge and adapt it to our structured output tasks.

Thirdly, we will explore open-source large language models such as LLaMA. These models, while smaller in size compared to proprietary LLMs, offer the flexibility of fine-tuning. By fine-tuning these open-source models, we can tailor them to our specific needs and evaluate their performance in comparison to other approaches.

Lastly, we will utilize ChatGPT-4o mini, specifically focusing on its prompting capabilities. Fine-tuning ChatGPT models can be money-intensive, so we will use it to predict data through well-crafted prompts. This method allows us to assess the quality of the generated data and its utility in training other models. In real-world applications, this approach can be a cost-effective way to create labeled datasets for training smaller, task-specific models.

By using these models, we aim to cover a broad spectrum of approaches, from simple baselines to advanced LLMs. This comprehensive evaluation will help us understand the trade-offs and benefits of each method, guiding us in selecting the most effective approach for structured output tasks in natural language processing.

5 General Reasoning

The core hypotheses of this study revolve around the comparative effectiveness of fine-tuning and prompting methods for structured output tasks in natural language processing. To test these hypotheses, we have selected a diverse set of models and datasets that will allow us to comprehensively evaluate the performance of each approach.

The datasets chosen for this study include multilabel classification tasks, which are representative of real-world scenarios where text needs to be categorized into multiple classes simultaneously. These tasks are particularly relevant for structured output applications, such as converting free-text into a structured format. By using these datasets, we can simulate the complexity and challenges of real-world structured output tasks.

The models we have selected range from simple baselines, such as logistic regression with TF-IDF features, to advanced transformer-based models like BERT and open-source LLMs such as LLaMA. Additionally, we will use ChatGPT for its prompting capabilities. This diverse set of models allows us to compare traditional fine-tuning methods with modern prompting techniques.

Fine-tuning BERT-like models on well-curated datasets will help us determine if this traditional approach consistently outperforms prompting methods. By fine-tuning open-source LLMs, we can evaluate the effectiveness of adapting smaller models to specific tasks. Using ChatGPT for prompting will allow us to assess the quality of generated labeled data and its utility in training other models.

The combination of these models and datasets will enable us to test our hypotheses in a controlled and systematic manner. We will measure the performance of each model using standard evaluation metrics for multilabel classification, with a particular focus on the exact match ratio. This metric is crucial for our task, as it requires the entire set of labels for each instance to be predicted correctly, reflecting the model's ability to fully understand and classify each instance.

By analyzing the results, we aim to identify the strengths and weaknesses of each approach, providing insights into the trade-offs between fine-tuning and prompting methods. This will inform our understanding of which techniques are most effective for structured output tasks and guide future research and practical applications in natural language processing.

6 Summary of Progress

Currently, we address the task of structured output—specifically, converting text to JSON—by writing detailed prompts. The resulting JSON is essentially a convenient representation of multiple nested multilabel classifications. The prompt includes comprehensive instructions on how the model's output should be structured, describes various exceptions, and provides examples of specific categories.

Our workflow with the prompt involves running data through it, manually reviewing the results, and refining the prompt to eliminate errors. Surprisingly, top commercial models handle this task well, and it is possible to craft a prompt that minimizes errors significantly. We have been operating under the assumption that a well-engineered prompt for a large model is always superior to fine-tuning open-source models. However, this hypothesis has not been rigorously tested.

The goal of this project is to test this hypothesis. We aim to determine whether it is possible to train an open-source model to achieve performance comparable to that of paid cloud-based LLMs. If successful, this could lead to significant cost savings by reducing reliance on expensive commercial models.

So far, we have:

- Experience in developing detailed prompts for converting text to JSON.
- Conducted tests with commercial models, achieving high accuracy on relevant tasks.
- Identified the need to compare this approach with fine-tuning open-source models.

What we still need to do:

- Select and preprocess appropriate datasets for training and evaluation.
- Fine-tune BERT-like models and open-source LLMs on these datasets.
- Conduct a comprehensive evaluation of the fine-tuned models versus the prompt-based approach using commercial models.

Potential obstacles and concerns:

- Ensuring the availability of high-quality, labeled datasets for fine-tuning.
- Balancing the computational resources required for fine-tuning large models.
- Accurately measuring and comparing the performance of different approaches, particularly in terms of exact match ratio.

By addressing these tasks and overcoming these challenges, we aim to validate our hypothesis and potentially identify a more cost-effective solution for structured output tasks in natural language processing.

References

[1] Kaggle. Sentiment Analysis for Mental Health *https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health*

[2] Kaggle. Multi-Label Classification Dataset *https://www.kaggle.com/datasets/shivanandmn/multilabel-classification-dataset*

[3] Huggingface. Mehdi Iraqi. Comparing the Performance of LLMs: A Deep Dive into Roberta, Llama 2, and Mistral for Disaster Tweets Analysis with Lora. *https://huggingface.co/blog/Lora-for-sequence-classification-with-Roberta-Llama-Mistral*

[4] Arousha Haghighian Roudsari, Jafar Afshar, Wookey Lee, Suan Lee. PatentNet: multi-label classification of patent documents using deep learning based language understanding. *https://link.springer.com/article/10.1007/s11192-021-04179-4*

[5] Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, Xiaoqin Zhong. Label Supervised LLaMA Finetuning. *https://arxiv.org/abs/2310.01208*